Tsila Hassine

Piet Zwart Institute

June 2006

Ctrl-F Reader: Some Questions On The Changes In Reading Practices And Knowledge Acquisition Processes Since The Appearance Of Digital Text Databases

Introduction

One of the main influences of electronic media on the digitized hemisphere is the abundance of content. Technology has made electronic content accessible, and recently also easily produced. The quantitative changes in access to textual content, and the qualitative changes in its storage format are bound to influence the way we consume it. In this essay I am reviewing some of the approaches that were initially developed out of the need to cope with these changes, and then later try to take advantage of this already established situation.

I start by describing the problem: my relationship with the web, and what drew me into this subject.

I then move on to describe different approaches that attempt to deal with this problem.

We can divide the different approaches described in this essay into 3 periods:

1: The early and relatively “naïve” period. Back then textual electronic databases were still new and they were intended for a professional public. The database was exploited using simple text searches that retrieved the lines around that text.

2: More advanced uses. The machine looks for the co occurrence of pairs of words and tries to interpret these co occurrences in a meaningful way

3: Critical tools: attempts to analyze electronic and textual content through frequency analysis.

4: Recent developments that try to mine the Internet in a way that will provide AI qualities.

The problem: too much text

The WWW is the most far reaching example of a human-machine hybrid. Human produced content is archived, tracked and organized by machines. Increased storage capacities, faster chips and broader bandwidth constantly blur the boundaries between the individual and the collectively built Internet machine. Customizable authoring platforms archive the daily musings of millions of contributors on mass storage servers. This giant archive of human content is kept accessible and usable thanks to the tireless efforts of thousands of spiders of different size and complexity. Those bots crawl the web on a daily basis feeding databases designed to serve a variety of purposes. Ranking algorithms determine who will be hit everyday by millions of eyeballs, and who will be hardly exposed even to the few surfers who managed to formulate a precise enough search query. Internet users' browsing, reading and viewing patterns are dissected and analyzed by sniffers and web stats, and their results affect the nature of future content.

It is interesting to compare the rates of content production and content consumption. Technorati claims to be tracking about 40 million blogs (1). Yet Nielsen//NetRatings estimated that in March 2006 the average number of domains visited per person per month was 71 (2). Blog aggregators enable the motivated user to read up to 1000 blog updates per day (not recommended for epileptics) (3). These differences are bound to enforce a powerful power law distribution (4) - where only a tiny fraction of the Blogs capture the attention of more than a few dozen eyeballs per month, while the most blogs (millions of them) are hardly read more than a couple of times. This quantitative difference naturally leads to the following question: who are bloggers writing for? Most blogs run the risk of ending up as "recommendation list fodder"(that is if they are lucky enough to be discovered by any of the aggregators), and merely serve as statistic figures for aggregators of all sorts, or for generating "Google Bomb"(5)-like phenomena (i.e tweaking search engine ranking through manipulative coordinated massive linking).

These incomparable numbers also raise the question of what we read more – plain text or statistics? And who manages to capture our attention? What are we reading more? Amazon users' book reviews, or Amazon's "people who bought this book …" recommendation list?

The conclusions from these numbers are almost self-evident: the individual user can't make good use of this abundance of content on her own. We are inherently unable to process this flux of information by ourselves. For lack of better content processing strategies we are “lost in data” (6) or we might even develop “Information Anxiety”(7) symptoms….

A possible approach to counter this situation would be to “fight fire with fire”. Machines help to push content down our throats? Let’s design machines that help us digest it. While still steering clear of the unsolved AI problems that may be involved in such an approach, there are ways to design machines to intervene in the reading/content consumption process without them doing the actual reading for us. Directories, search engines, blog aggregators and RSS feeds automate selection and ranking processes. The simple “ctrl-f” command has also proven to be a useful hack to the unmotivated reader when required to fish the proverbial data needle from the textual haystack.

But the richness of available content-, and the abundance of content-creation and collaborative platforms, allow – even call – for more personalized ways to navigate this ocean. Some artistic projects (which I review below) try to address this issue, designing new reading approaches and suggesting strategies to facilitate excessive content consumption.

I too had to find my “customized” approach to this medium, as its role in my daily routine kept on expanding. On countless occasions, I found myself at loss, trying to handle an unreadable, unreasonable mass of content, being able to absorb only a tiny fragment of the whole. The traditional way of reading - obediently following the lines top to bottom, left to right -- didn't get me far in this hypertextual galaxy. Necessity being the mother of invention, and recognizing my inability to beat the machine on its own text-processing turf, I decided to join forces with it, and harness it to my own needs. The search engine has since become my ally, and ctrl-f became my second pair of eyes. I have come to realize that reading is turning into a chain of decisions, each eliminating large quantities of text. The ctrl-f reading part of this work tries to systematize this ctrl-f phenomenon, and thus to acknowledge both the widespread use of this practice, and the decisions it helps the reader make.

Putting the machine to work:

Here I describe how the degree of involvement of the machine in the practice of reading evolved since the appearance of early electronic textual databases. I start with simple "ctrl-f" like approach used to mine the first databases. Then I describe a more recent approach were the machines tracks co occurrence of words. I add a short review of how statistical approaches are used in order to analyze content. Finally I describe 2 very recent projects that attempt to use the WWW for AI purposes.

The early period: advent of the digitized text corpora

In order to try and understand the ways in which the giant hypertext repository called WWW changes and will continue to change our practices of reading and processing text, it is perhaps worth while to examine test-cases of earlier digitizations of large text corpora in specific areas, and examine how these corpora changed practices of studying, processing and analyzing information. In this section I describe the results presented in a paper (8) that presents such an investigation, and discuss some of the issues that arise from this transition. I will also discuss some of the similarities and differences between the textual database presented in the paper and the WWW.

In an early paper (8) "'Pulling down' books vs. 'pulling up' files: textual databanks and the changing culture of classical scholarship"(1995) Karen Ruhleder investigated the ways in which the integration of a textual electronic databank into Classic studies influenced study and research practices. This electronic textual databank - the "Thesaurus Lingua Greca" (TLG) contains electronic versions of ancient Greek literature, and other materials such as Latin texts and other Greek materials. In addition to these texts the TLG also provides a strong infra structure for searching and processing these materials. The researcher based the paper on the result of 60 interviews with faculty members in Classic Studies departments, tool developers and editors of traditional journals and electronic mailing lists, in an attempt to investigate possible changes in certain reading practices due to the integration of large electronic text repositories.

The fundamental principles manifested in this transition from scattered printed material (Greek manuscripts, concordances etc) to a unified digitized hypertext system equally available to all are also a characteristic of the WWW, and therefore relevant as a case-study for the investigations of how the WWW and its principles of accessibility, hypertextuality and abundance may affect the general public's approach towards information consumption and processing.

In Classics studies texts are a key resource, serving both as a source for discoveries, and as source of support for evidences. An example for a work process is that upon "discovery" of a word in a certain text, the researcher needs to look for other occurrences of the same word in other texts. For that purpose the researcher uses dictionaries, indexes and concordances to see where and how that word was previously used. It's a laborious process and sometimes problematic since not all books are available at all times. Under such conditions, a researcher often has to rely on intuition to guide her through that process. Such intuition – a "feeling" towards printed matter – develops after years of leafing through the same books in quest for different answers.

The TLG was originally conceived as a databank for Greek literary texts. It currently contains electronic versions "of all of ancient Greek literature surviving from the classical and post-classical periods" (Brunner 1987)(9). Some earlier attempts to encode selected texts (mainly the Homeric texts) were made in the 1960's by different individuals, but none had the same impact as the TLG. To use the TLG the scholar enters the word or phrase he is searching for, and selects a search range (a single text, a specific set of texts, or all the texts). When the TLG finds a match, it presents the text within context: the line in which the search phrase occurs, and the lines that precede and succeed it.

The immediate benefit of such a computerized repository of all texts is their constant online availability. This provides researchers with a richer sense of the breadth of available material. As the author puts it: "with the TLG …texts that once lay at a scholar's mental periphery are pushed center stage on to the computer screen" (10). This repository and its search options present some further advantages. Computerized searches are more accurate than the human eye, or poorly constructed concordances. They allow the construction of questions of wider range -- since crossing the interdisciplinary divide is made easy. A researcher can make statements about series of texts without having actually read them, just by automatically searching through them. This changes the traditional relationship between scholar and text. According to this paper "knowing" a text is replaced by knowing how to construct a search algorithm. This possibility reduces researchers' willingness to read texts in their entirety: "[Without the TLG] I would have to read…through the orators, which is a pretty large corpus, and the philosophers, which is huge"(11). On the other hand some researchers find that using the computer "detaches you from your work…sometimes you want to completely digest, master, live with the text". Processing texts in such a fashion that detaches them from their broader context may affect the researcher's abilities to produce in-depth contributions about the texts they are researching, and to understand the value of their findings within the larger body of scholar work.

Some scholars mention also the disappearance of the serendipitous aspect: "When you read, you learn things that you aren't even searching for. The human brain may be searching for X, but will be confronted with other things"(12).

These reflections raise not only the question of how a good research can be practiced upon material mostly taken out of context, but also at an earlier stage in the research process – the stage when research questions are formulated, and research paths are outlined. Research is a dynamic process. It evolves from one question to another. At each stage the researcher asks a question, searches for answers, and formulates the next question based on these obtained results. When research is text-based, as in the case of Classic studies, does the existence of the TLG influence the nature of the “next question”? Is the research path shaped by the machine’s ability to parse text? Since the TLG saves researchers the trouble of reading numerous books in quest for answers, maybe it can be taken one step further in saving researchers formulating a set of questions. A possible “research algorithm” could be the following:

```
For question Q1 – search for all occurrences of string S.
        If the obtained set of results has property P1 – perform search for question Q1_2.
        If the obtained set of results has property P2 – perform search for question Q1_3.
.
.
.
```

(It may be an interesting exercise then to try and divide research questions into “machine answerable” questions, and “non-machine answerable” questions….)

Another issue is the confidence level researchers have in such a tool. What part of the research does this tool occupy? Does the researcher use this tool as a mere reference tool, sort of a pointer towards what books will be likely to produce the desired answers, or does the researcher rely solely on the results provided by the repository and its search algorithms? Moreover – does the researcher formulate questions based upon a prior

understanding of the machine's abilities? Or in other words – is there a real risk of research turning into "TLG oriented" research? As the author puts it, "will this 'technological' character translate into the growth of statistically based research?"
I believe now is a good time to ask such questions, since digitized text corpora are just now becoming more and more available, and traditional research practices are at the risk of becoming obsolete.

Another issue is the question of the editorial processes – the accumulation of thought constructed in the hundreds of years where books were still the only reference. In the specific case of the TLG this has been left out, but does it necessarily have to?
Also, the inclusion/exclusion process -- being able to differentiate between -what is important and what is not- is a fundamental process that involves experience with the matter in question, intelligence, and the process maker's own voice. In the case of the TLG, the "voice of the community" was lost (13).

In Classic Studies, there is an important element of the traditional textual edition and that is the "apparatus criticus". This is the set of textual footnotes that point to previous works. This part is physically present along with the text and serves as a reminder that the text is one person's "best guess" at the moment. As one scholar explains this: "Creating an apparatus is a very labor-intensive and intelligence-intensive process….Because of the way the apparatus is written in relation to the text that's above it, it takes some intelligence to decide what to put in your version…you have to make a decision…"(14)

The TLG doesn’t include any critical notes, and only offers one version of each text. This lack of critical materials deprives the scholar of an important set of links to other scholars, both past and present. Textual editions include information about the text, and its history throughout generations of Classic Studies. The decision not to include this part in the TLG doesn’t necessarily derive directly from the electronic nature of the medium. It may be due to copyright laws, or to the TLG creators’ decision to provide mainly a repository for ancient texts. Therefore this decision can be changed in a later version of the TLG. But when approaching a project of this magnitude, the editors shouldn’t aspire only to provide the richest repository of original texts and ignore the body of work that was produced around them for hundreds of years. Since electronic space is practically limitless, careful consideration should be given to the issues of how to represent not only the original texts but also the richness of research it has triggered.

While the TLG isn’t as broad as the WWW, its underlying principle (computerized textual searchable database) produces some similar effects. But we also see that different design decisions play an important role.

The issue of constructing good search algorithms is very reminiscent of the practice of looking for something on the web. Constructing good search queries has become an important skill, as the user wishes to produce a small and accurate set of results to avoid spending time going over too much text. In the case of the Internet, computerized searches not only produce more accurate results, but without them searching the web is impossible. On the other hand, the aforementioned serendipity component is still present on the WWW, though missing in the TLG. The WWW user often “gets lost” in the Internet. Searching for one thing very frequently leads you to findings of a totally different nature. This is made possible on the WWW through its hyperlinked structure. This structure wasn’t implemented in the TLG, and may be one of the reasons for the disappearance of “serendipity” from the TLG, as the researchers have indeed complained. Where the WWW does differ from the TLG is in the issue of openness and bias. The WWW is an (relatively) open system, while the TLG is a closed one. This means that anybody with Internet access and medium level computer skills can publish their own web content. The TLG, on the other hand, was produced by a small group of experts who decided to turn it into a repository of all original texts. On the other hand, where the TLG abolishes bias that comes from accessibility difficulties, the Internet creates one. Whereas researchers no longer depend on library and travel grants to access remote material, Internet access creates a gap between those who have easy Internet access and also possess the required skills, and those who don’t.

Ruhleder’s paper shows us how digitized text corpora affect out relationship to text and our practices of text processing. We also see that abundance of space can affect editorial decisions, and what is the influence of database design decisions. Finally we see points of similarity and difference between such a system and WWW.

Intermediate period: counting word co-occurrences for linguistic information

In this part I will describe more advanced uses of large electronic text corpora. The approaches I am describing here differ from the simple text extraction described above, in this case another parameter is sought: frequently co occurring pairs of words. In this way researchers claim machine is able to "learn" human language through statistical processing.

Extracting Language Based Knowledge From Large Digitized Text Corpora

In a paper from 2002 titled "The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches" by Reinhard Rapp (14), the author describes the use of statistical methods in automating the performance of language related activities. The paper shows that the two types of relations as defined by de Saussure are reflected in the distribution of words in large text corpora.

According to de Saussure, there are two fundamental types of relations between words: the syntagmatic and the paradigmatic. There is a syntagmatic relation between two words if they co-occur in spoken or written language more frequently than chance would have it, and if they have different grammatical roles in the sentences in which they occur. Typical examples are "coffee – drink", "teacher – school". A paradigmatic relation between two words occurs if the words can substitute for one another in a sentence without affecting the grammaticality or acceptability of the sentence. Typical examples are synonyms or antonyms such as "quick – fast" or "eat – drink".

It is possible to automatically detect paradigmatic associations. Paradigmatic associations are words with high semantic similarity. According to Ruge (16), the semantic similarity of 2 words can be computed by comparing their lexical neighborhoods. For example, the semantic similarity between “red” and “blue” can be derived from the relatively high frequency of their co-occurrence with similar word sets such as “color”, “flower”, “dress”, “car” etc. Counting the co occurrences of word pairs is performed by counting how often this pair occurs within a window of text of fixed size. This simulation is based on regularities in the statistical distribution of words in a large text corpus. In this case the corpus is the British National Corpus (BNC), a 100 million-word corpus of written and spoken language compiled with the intention of providing a representative sample of British English.

In order to measure the success of this method, the system was tested against the synonyms part of the computerized version of the TOEFL (Test of English as a Foreign Language) test: The computer was given the stimulus word and compared its co-occurrence vector against the co occurrence vectors of the proposed synonyms. The results obtained with this method ranged around a 69% success rate (the human rate is 64.5%). The differences in results are due to the different implementation of the algorithm and the use of different optimization methods.

It is also possible to detect syntagnatic relations. Syntagmatic associations are words that frequently occur together. Therefore, the obvious way to recognize a pair of words that respond to that definition is enough to look for word pairs whose co-occurrence is

significantly higher than chance. To compute the syntagmatic associations with a stimulus word we need to find the co-occurrences of other words with this specific one. Computationally speaking, this is of course significantly simpler than extracting paradigmatic relation where co-occurrence vectors have to be computed for all words. As in the paradigmatic case, co-occurrences are counted
within text windows of fixed size.

Extracting Language Based Knowledge From the WWW

In another paper from 2002 titled: "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL"(17), researcher Peter Turney brings an innovative approach to synonym discovery by using the Internet as its text corpus.

The author describes a new method he developed for synonym recognition. Given a "problem" word and four alternative words, the algorithm has to indicate which of the alternatives is most similar in meaning to the problem word. For each alternative, the algorithm asks the Altavista search engine how many documents contain both problem word and the alternative word. This number is then divided by the number of documents that contain only the alternative word. This ratio is the score for the alternative word, such that the alternative word with the highest score is considered most likely to be synonymous with the problem word. This method, seemingly the conceptually simplest of all methods for synonym discovery, produces the highest scores on the TOEFL test, a

74% success rate as opposed to 69% in the previous paper. In this case the hard work isn’t done by the algorithm, but by AltaVista.

A possible reason for the success of this seemingly simple strategy can be found in the following quote from Landauer and Dumais (18) when discussing their own algorithm (which didn’t involve the Internet):”We know of no other fully automatic application of a knowledge acquisition and representation model, one that does not depend on knowledge being entered by a human but only on its acquisition from the kinds of experience on which a human relies, …”. This is where the Internet factor creates the difference. Humans have come to rely heavily upon the Internet in the recent years, and the Internet’s participating audience has grown and diversified (to some extent). The easy and collaborative authoring platforms such as Wikipedia and Myspace have drawn a more diversified group of contributors, and also provided motivation to contribute on a wide range of topics. Therefore, Internet language reflects the use of language among a wider part of population. The BNC or other text corpora are limited when compared to the Internet, for two reasons. The first is that they were compiled by a small group of researchers, while the texts found on the Internet were created by millions of users. The second is the time factor: the Internet is constantly changing, new texts are added every minute, and therefore Internet language is constantly adapted to the spoken language and cultural changes. This is impossible to achieve in any text corpora. Therefore methods exploring the use of modern language will benefit from using the Internet as their text corpus.

An important issue that wasn’t addressed at all in these last two papers (or in other papers concerned with computational linguistics) is the nature of the sample upon which these linguistic machines feed, and the possible effect this has on the obtained results. In both cases the algorithms process information provided by digitized text corpora that reflect the use of the English language among residents of a specific country (such as the BNC) or among population with regular internet access, with basic computer skills and relatively well versed in Internet culture. Results obtained from processing these samples can hardly reflect any possible “universal” use of language, and are limited to the language used by the persons who have contributed to that sample. We should also note the benchmark against which the quality of these results was tested: the TOEFL test, which is meant to test the level of English proficiency among candidates applying to universities in English speaking countries.

One could argue against this critique and claim that these algorithms do not pretend to derive any broad conclusions regarding any “universal” nature of language. They only test the ability to systematically process texts and derive certain conclusions from it. The results should be regarded as an evaluation of the machine’s success in learning language fed to it in a certain format. We should always be aware that the only “intelligence” a machine has is limited to the set of instructions upon which it operates. In these two examples the machine had very simple instructions (retrieve words from a limited window of text), and therefore should not be expected to produce anything more than statistical estimates of word co occurrences in different text corpora be it BNC, Internet, or SMS Hebrew.

From these two papers we see that applying co occurrence and frequency counting methods on the Internet text corpus leads to good results in predicting use of words in certain human languages. In my own work I expand on his principle a little further, and repeat the same methods using as stimulus words names of personalities or issues. In this way I hope to predict the way these issues are referred to over the Internet. By narrowing my search to specific content providers I hope to discover whether their use of language differs from one another when referring to specific issues. I wish to find out whether there is a certain “CNN on Bush” speech, as opposed to “Bush in the Blogs” speech (my initial hunch was that “blog speech” would be at least more diversified). I am trying to chose sources that are regarded as influential, and that are supposed to take different stands regarding the same issues. I am also trying to integrate the time factor in this work, (something which wasn’t done in the papers described above). By tracking the frequency and co-occurrence of words within these specific sources over time, I wish to see whether these content providers’ use of words changes over time as events happen, and how such changes take place.

Developing machine critique?

In this part I will take a little break from digitized text in order to take a look at the way the abundance of content transmitted over electronic channels has affected Communication studies, and what statistical critical tools were developed around it.

Counting event frequency and event co occurrence as a critical tool

The 1960's saw the emergence of empirical ways of approaching communication, meaning and sign – areas usually covered by semiotics. Researchers in communications studies started to develop experimental methods to test and prove (or disprove) the reliability of their hypotheses. One such approach is content analysis.

"Content analysis is designed to produce an objective, measurable, verifiable account of the manifest content of the message…. It works best on a large scale: the more it has to deal with, the more accurate it is. It works through identifying and counting chosen units in a communication system. …The units counted can be anything that the researcher wishes to investigate: the only criteria are that they should be readily identifiable and that they should occur frequently enough for statistical methods of analysis to be valid."(19)

Words, naturally enough, are often counting material. In an example from 1967, Paisley (20) analyzed the four televised debates between Nixon and Kennedy. He counted the number of times they used particular words during those debates, discovering differences in the frequency with which they used the words: "treaty", "attack" and "war".

Table1: *Kennedy and Nixon: word frequency. Frequency of use per 2500 words.* (21)

Word	Kennedy	Nixon
Treaty	14	4
Attack	6	12
War	12	18

This data can be used to provide some evidence of the candidates' different attitudes: they could be interpreted as if Nixon was more bellicose, while Kennedy was more conciliatory. But then such difference can also be due to different reasons. This illustrates the main problems with such an approach, problems that result when words are "ripped" out of their contextual environment.

To practice content analysis, one must adhere to certain requirements. It must cover the whole message or message system, or have a properly constituted sample. If it claims to have some scientific objectivity, it can't select particular parts of a message while ignoring others.

Content analysis can provide a tool to test a particular "feel" towards the subjective, selective way in which we receive messages. One example was an examination of job stereotypes in American media in the 1970's. Seggar and Wheeler (1970) (22) studied women' s job stereotypes in American television (Table 2).

Table2: *Five most frequently portrayed occupations on American television according to race and sex* (23)

Males occupation	%	Females occupation	%
		Blacks	
(N=95) Govt. diplomat	 18.9	(N= 20) Nurse	 30.0

Musician	13.7	Stage/Dancer	15.0
Policeman	9.5	Musician	5.0
Guard	9.5	Govt. Diplomat	5.0
Serviceman	5.3	Lawyer	5.0
Total	56.9	Secretary	5.0
		Total	65.0
British			
(N=104)		(N=17)	
Guard	13.5	Nurse	41.2
Musician	11.5	Secretary	11.8
Waiter	7.7	Maid	5.9
Physician	4.8	Govt. Diplomat	5.9
Serviceman	4.8	Actress	5.9
Total	42.3	Total	70.7
White Americans			
(N= 1,112)		(N=260)	
Physician	7.6	Secretary	15.4
Policeman	7.6	Nurse	15.0
Musician	4.8	Stage/Dancer	8.1
Serviceman	4.6	Maid	6.5
Govt. Diplomat	4.5	Model	5.0
Total	29.1	Total	50.0

Note: N=actual numbers in samples

Content analysis was also systematically applied to the news. A series of strikes and their TV coverage were the focus for the Glasgow Media Group (a research based grouping of academics within the sociology department of Glasgow University) in the 70's. Their work consisted of analyzing television coverage of industrial news. They investigated coverage of strikes in different sectors. They discovered that television coverage of three sectors - motor industry, transportation, and public administration – was substantially higher than printed press coverage. This was also the case with other parameters such as total number of work days lost, total number of workers involved, or number of stoppages as recorded by the Department of Employment statistics. Content analysis revealed that this pattern of distorted reporting didn't reflect patterns in reality. But this analysis couldn't answer the question "why?" Content analysis can reveal *patterns*, but can't always discover what *causes* them.

George Gerbner, founder of the Cultural Indicators project, is one of the leading figures in this area. He wasn't content with merely revealing patterns, but endeavored to discover their causes. He developed a theory of how content analysis can explain culture and its impact. He is perhaps best known as "The man who counts the killings", after famously estimating "that the average American child will have watched 8,000 murders on television by the age of 12" (1970) (25) (this estimate was reported in the 70's, and I believe is still relevant).

Gerbner believed that culture communicates with itself through its total mass-media output.

For Gerbner, the great strength of content analysis is that it analyzes the whole message system. It is the "massness" of the media available to the individual that is significant, not the selective individual experience of it. And it is exactly with this "massness" that content analysis deals best. Content analysis reveals the patterns that lie under the whole output. These patterns are the important characteristics of the media.

Content analysis can and does reveal patterns and frequencies in the denotative order of communication. These patterns and frequencies reveal values and attitudes.

The studies by Gerbner (1970), Seggar and Wheeler (1973), and Dominick and Rauch (1972) suggest some conclusions regarding dominant social values. Over-representation of men, white-collar jobs and certain races lead to the conclusion that frequency of such portrayal is related to a high ranking in the conventional value system.

Yet this approach suffers from a fundamental flaw. This statistical/ analytical approach is the results of the abundance of content, which required for statistical tools. Yet those same statistical tools tweak the content and rips the surveyed events out of their original context, thus leaving them open to interpretation that might differ from the original message. This problematic can is wittily summarized in the German saying: “never trust statistics you haven’t tweaked yourself” (26).

The works described here mainly involve counting word frequencies over varying periods and across different channels. But described above, looking at co- occurrence of words may also be a valuable *critical* tool.

Framing – a recent model developed by linguist George Lakoff (27) explores the language used in reference to certain issues, or the words used to “frame” them. As Lakoff explains in the introduction to his book, ”You can’t see or hear frames. They are part of what cognitive scientists call the 'cognitive unconscious' – structures in our brains that we cannot consciously access….We also know frames through language. All words are defined relative to conceptual frames….Every word… evokes a frame, which can be an image or other kinds of knowledge…When we negate a frame, we [still] evoke that frame”. (28)

As an example, he points out that when George W. Bush went into office, the combination “tax relief” started to be heard quite frequently, thus insinuating that taxes are an affliction, which George Bush will attempt to remove (29).

With the same methods used for synonym extraction, it can be relatively easy to track such attempts of such issue framing, especially since important parts of textual media are in digital format.

Since the Web has become such an important communication channel, it should be interesting to apply similar analysis on the content it publishes. Moreover since the web consists mainly of text in electronic format it lends itself easily to analysis that consists of detecting patterns and counting repetitions, which are the primary tools for content analysis. Therefore analysis of web content seems extremely simple, and should be made openly available to every Internet user.

The aim of a tool to automate web content analysis would not be to automatically generate a mass of research papers that analyze the frequency of use of this word or that, but rather to provide the regular media consumer with a simple tool to help scrutinize the nature of the media she consumes.

Every media consumer should be able to easily test whatever particular “hunches” she may have regarding the messages she receives from digitized media. The availability of such a tool should encourage the user to develop a growing awareness of the possible ways in which the user’s favorite media channel may be biased. Moreover, the ease of use of such a tool should also emphasize the ease with which electronic digital media lends itself to automatic processing and manipulation, and hopefully encourage the user to devise more analysis-oriented tools.

Of course such a tool also has problematic aspects. Users untrained in surveying methods are very likely to use it in ways that might lead them to biased conclusions. Moreover, in view of the wealth of content available on the web, it might be quite challenging to form properly constituted samples, and therefore provide significant results. Also, we can’t ignore that the internet itself is a biased medium, where English is the *lingua franca*, and which provides content mainly produced for and consumed by users with regular Internet access, basic computer skills and a certain income.

Taking these problems into account, I still believe such a “personal” analysis tool is an important companion to media consumers. While this tool may enable the user to derive conclusions about the whole of the Internet, or may not be a meaningful addition to those without regular Internet access, it will still enable the regular Internet consumer to exercise some critique and have a slightly higher degree of control over the media he consumes, helping him examine the mediated entourage in which he lives.

Making the machine learn

In this part I look at two recent scientific projects that attempt to make the machine produce "real knowledge" based on material harvested from the Internet, based on its textual content. This approach is significantly more advanced than the previous approaches described earlier in its ambitions, in the methods it employs and in the degree of autonomy it allows the machine.

The Internet (And Its Search Engines) As A "Common Knowledge" Base

For several years it seemed that AI was stuck, that it had reached its limits. As powerful and sophisticated computers have become, they still lacked the intelligence of a 3-year old. On the surface machines appear to be getting more intelligent all the time: robots mow lawns, printers and washing machines diagnose their own problems, and even car engines seem to have a silicon brain of their own. Yet none of these machines, or even state of the art computers can hold a simple conversation with a human being for more than 3 minutes. This is known as the *Turing test* and has been deemed extremely difficult to overcome. Yet it seems that a conceptual breakthrough may enable the machine to learn some common sense on its own. This breakthrough consists mainly in recognizing the value that human input may add to the common sense machine. Luckily enough, the

Internet may be just the thing AI experts were waiting for. In the last 10 years the Web emerged as a huge repository of electronically available knowledge, and indexing systems such as Google have made that knowledge easily accessible, both for human and machine.

As Michael Witbrock from the Cyc project in Austin phrased it: “The web might make all the difference in whether we make an artificial intelligence or not,”(30).

The Cyc project

Cyc, initiated by AI veteran Douglas Lenat, was conceived with the aim of replicating the human brain’s reasoning and learning ability. It relies on the idea that effective machine learning depends on having a core of knowledge for newly learned information - known as “common sense”. For 20 years a large enough core of knowledge has been entered into this project, and Cyc scientists developed a method of using a combination of Cyc and the WWW (accessed via Google) to assist in entering knowledge into Cyc. The process combines information gathered from the web with the existing Cyc knowledge base. I n their paper (2005) (31) they describe the process through 6 main steps, illustrated by an example which (presumably) unintentionally also serves as an excellent illustration of the highly controversial nature of their novel epistemological approach,

The process:

1. Selecting a query: an interesting query is created through an automated process that selects several parts of a sentence from the Cyc knowledge base. These parts

are composed together into a sentence that is judged to be the most interesting and productive, i.e that is the most likely to be found by a web search.

An example of such a query might be:

(foundingAgent PalestineislamicJihad ? WHO)

2. The chosen sentence is fed as a search string into Google's API. The query from above may be rendered into one or more "English search strings such as:

   "PIJ, founded by"

   "Palestine Islamic Jihad founder"

   The resulting documents are retrieved and the relevant sections extracted.

3. Once a document is retrieved, the answer must be found and interpreted. The system searches for the exact search string, and returns the search string plus either the beginning or end of the sentence. It is then parsed into the Cyc base and results in a string that looks like this:

   (foundingAgent PalestineIslamicJihad Terrorist-Nafi)

4. The resulting string is searched for phrases that can be interpreted by Cyc knowledge base rules. It is then checked against the Cyc knowledge base for consistency with already known facts.

5. In order to guard against parsing and excessively general facts, a second Google search is performed in order to determine whether the newly generated Cyc fact

will produce results. The search queries taken from sentences obtained in stage 3 (and verified in 4) look like this:

"Bashir Nafi is a founder of Palestine Islamic Jihad"

A fact for which no results were produced is considered "unverified", and will not be presented to the review of stage 6.

6. In this final step, new facts are reviewed by a human curator, and if correct, are inserted into the Cyc knowledge base.

The original goal of Cyc was to provide a knowledgebase broad enough to support learning from language the way humans are capable of. The widespread popularity of the electronically accessible Web together with its efficient indexing systems may make it possible for machines to learn from language. Yet their approach raises many questions regarding the viability of the results obtained in such an unsupervised manner, and is discussed in some more length further below.

NGD – what can be learned from Google? – searching Google for meaning.

Cyc are not alone in their quest of producing AI based on WWW and Google. I am presenting here is another approach, somewhat less ambitious, which does not require implementation of sophisticated learning algorithms or extensive knowledge bases. It consists of some simple calculations based on Google's result numbers, and yet provides another way of "reading" the web.

Paul Vitanyi and Rudi Cilibrasi from the CWI in Amsterdam claim they found a method for automatically extracting the meaning of words and phrases from the WWW using Google results' numbers (32). Using their method the claim computers can learn the meaning of words by “plugging” into Google, and automatically extract the meaning of the words.

Their method is (relatively) simple, and its success is due to the fact that the WWW is the largest database on Earth, and the content produced by millions of independent users averages out to provide automatic meaning of useful quality.

Their approach is based on the realization that a Google search can be used to measure how closely two words relate to each other, and is very similar to the method developed by Peter Turney (2002)(17) for automatic discovery of synonyms (described earlier in this essay).

Given a pair of words, three numerical values are retrieved from Google: the number of results for each of the words independently of each other, and the number of results for a query composed of the pair. These results are used in a formula that normalizes the results against each other and against the overall estimated size of the indexed WWW. The resulting numbers thus comply with the definition of a metric space, and are acceptable as a legitimate “Google distance” between the two words.

Using this technique the authors demonstrate its ability to distinguish between colors and numbers, and between 17th century Dutch painters. In one of the examples provided in the paper, the names of fifteen paintings by Steen, Rembrandt, and Bol were entered. The names of the associated painters were not included in the input. The output provides a separation of these paintings according to the painters.

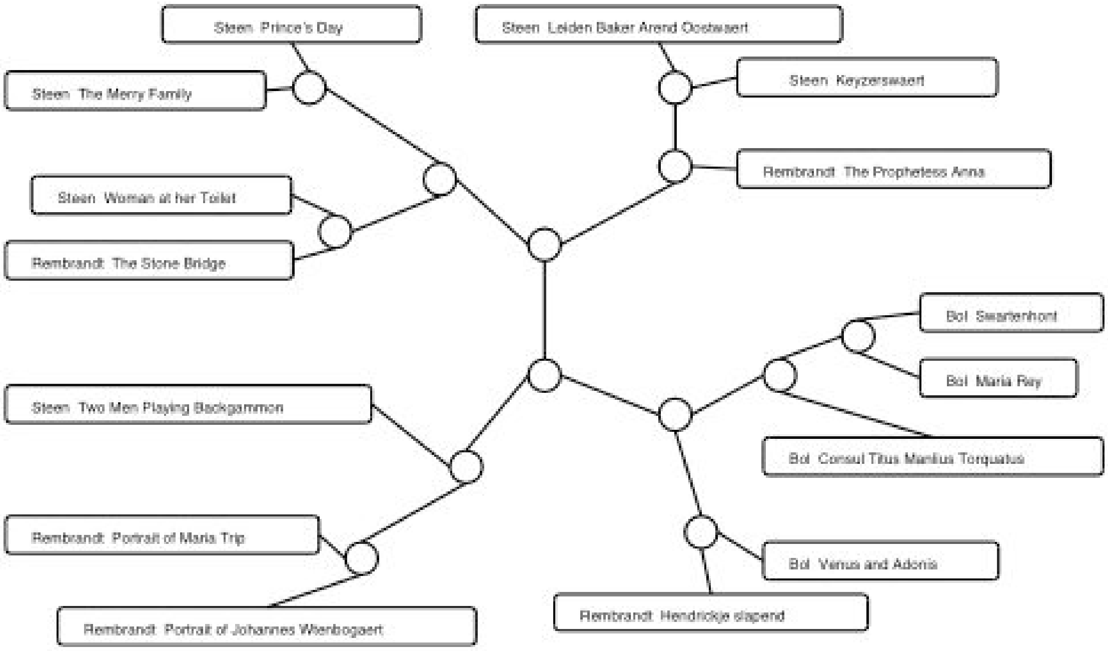

Steen Prince's Day
Steen Leiden Baker Arend Oostwaert
Steen The Merry Family
Steen Keyzerswaert
Rembrandt The Prophetess Anna
Steen Woman at her Toilet
Rembrandt The Stone Bridge
Bol Swartenhont
Bol Maria Rey
Steen Two Men Playing Backgammon
Bol Consul Titus Manlius Torquatus
Rembrandt Portrait of Maria Trip
Bol Venus and Adonis
Rembrandt Hendrickje slapend
Rembrandt Portrait of Johannes Wtenbogaert

Figure 1: Fifteen paintings tree by three different painters arranged into a tree hierarchical clustering. In the experiment, only painting title names were used; the painter prefix shown in the diagram above was added afterwards as annotation to assist in interpretation. The painters and paintings used follow. Rembrandt van Rijn: Hendrickje slapend; Portrait of Maria Trip; Portrait of Johannes Wtenbogaert ; The Stone Bridge ; The Prophetess Anna ; Jan Steen: Leiden Baker Arend Oostwaert ; Keyzerswaert ; Two Men Playing Backgammon ; Woman at her Toilet ; Prince's Day ; The Merry Family ; Ferdinand Bol: Maria Rey ; Consul Titus Manlius Torquatus ; Swartenhont ; Venus and Adonis (33)

While the effort to create intelligent machines is certainly laudable, and acknowledging the critical contribution human input might add to it, the approach presented in these two papers, and especially the amount of trust expressed in the internet as the modern day "*bocca della verita*" is quite alarming, especially in the way its used in the Cyc project.

In the abstract of the Cyc paper, the authors claim that their “long-term goal is automating the process of building a consistent, formalized representation of the world in the Cyc knowledge base via machine learning”. The problematic implications of this formalized web-based epistemology are practically crying out of this sentence and throughout the paper. Whose world? Whose world is going represented in the Cyc knowledge base? Whose world is entitled to be represented on the Internet? How much trust can be put in “knowledge” extracted from the Internet – and therefore how trustworthy could a knowledge base based on parsed internet content ever be?
The example provided in the paper to illustrate the knowledge-building system is also a perfect illustration of these problems. Asking the web to identify the leader of an organization called “Palestinian Islamic Jihad”, and then storing this in the knowledge base that is supposed to provide a representation of the world is very controversial. Such “evidence” isn't accepted in any institution. For all I know, the aforementioned organization can be a cartoon gang, and their leader a cartoon cult figure – of course its common knowledge then that “Bashir Nufi” is the leader of “Islamic Palestinitan jihad “.

Even if we accept the Internet as a reliable knowledge base, there is another problem – their faith in Google. It is common web knowledge that Google indexes only a fraction of all Internet pages, and that it has its own indexing agenda, so again we are faced with the question: whose world will be represented in the Cyc knowledge base? Only the Google-accredited world? If in Google we trust, then according to that methodology the famous “Bush is a miserable failure”-Google bomb will also find its respectable place as part of the Cyc consistent and formal representation of the world.

While the NGD paper also refers to the Internet and its Google representation as a knowledge base, but their approach is less pretentious, and they don't aspire to global representation. The examples they provide in the paper are not so politically charged (number names, colors and Dutch 17th century painters), and they acknowledge in a certain way the epistemological limitations of their approach. The name they chose for their model – "normalized Google distances" – reflects their awareness of the fact that the obtained results are determined by Google (and therefore also by the web). Moreover, the application they developed is available online for immediate use together with a simple interface that enables every user to run their own experiments and test their own hunches (34).

In these scientific papers we see that the web has developed in such a way that it can be used as a way to probe the internet users' thoughts (or at least parts of them). These two papers represent approaches that while they are similar in their underlying principle of Google usage they are different in the amount of trust they chose to put into that approach, and also in the possible application they wish to develop from it.

Artistic approaches

There are several projects that attempt to deal with the abundance of digital text. I will focus on a few notable ones.

Schoenerwissen in collaboration with Prof. Hans Ulrich Reck - txtkit: reading as a collective practice.

In this work, the Schoenerwissen duo refer to reading as a collection of decisions. When a person reads a text, she decides whether to read a certain paragraph based upon the first line of that paragraph. Therefore they provide a platform that enables taking such decisions in a collaborative fashion. It is an "Amazon" like platform, where the user is presented with the information that "this is the number of persons who read that first line and chose to read the whole paragraph"…The visual representation consists of elliptic curves that represent the different stages of collaboration. (for some reasons Schoenerwissen's site was offline by the submission deadline, therefore I am unable to provide any visuals)

http://www.txtkit.sw.ofcd.com/

Ben Fry – Valence: text as organic, self-organizing form.

Valence is an application that builds visual representations of very large sets of information. The author is interested in finidng models and representations for dynamic sources of data. Fry employs methods from behavioral studies and distributed systems where every piece of information is treated as an element in an environment that produces a representations based on their interaction – for example in large texts – the more frequently used words make their ways outside of the cluster and are thus more visible.

The author claims that the best way to transmit an immediate understanding of a large body of information is by providing a "feel " of how the information is structured.

http://acg.media.mit.edu/people/fry/valence/

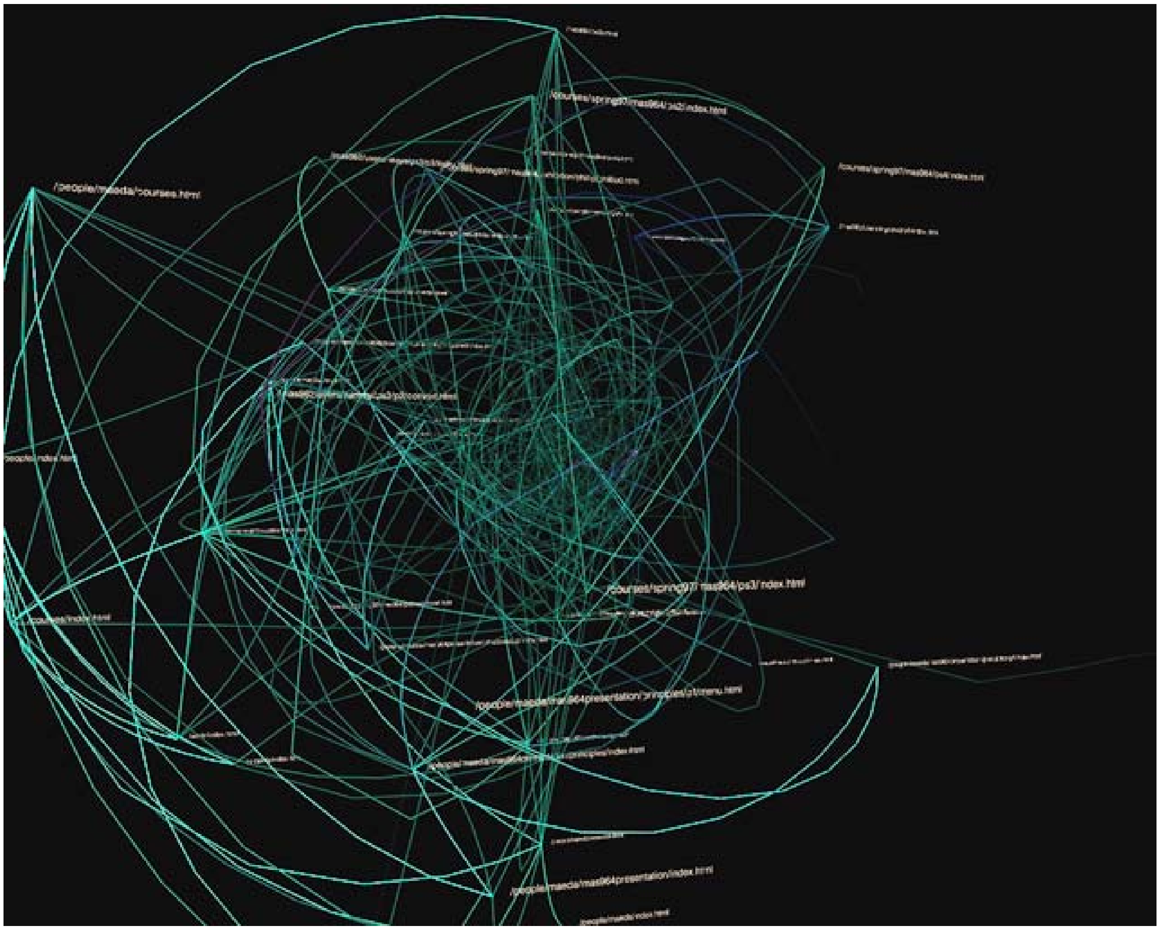

/people/maeda/courses.html
/courses/spring97/mas964/ps2/index.html
/courses/spring97/mas964/ps4/index.html
/people/index.html
/courses/index.html
/courses/spring97/mas964/ps3/index.html
/people/maeda/mas964presentation/principles/p1/menu.html
/people/maeda/mas964presentation/index.html
/people/maeda/index.html

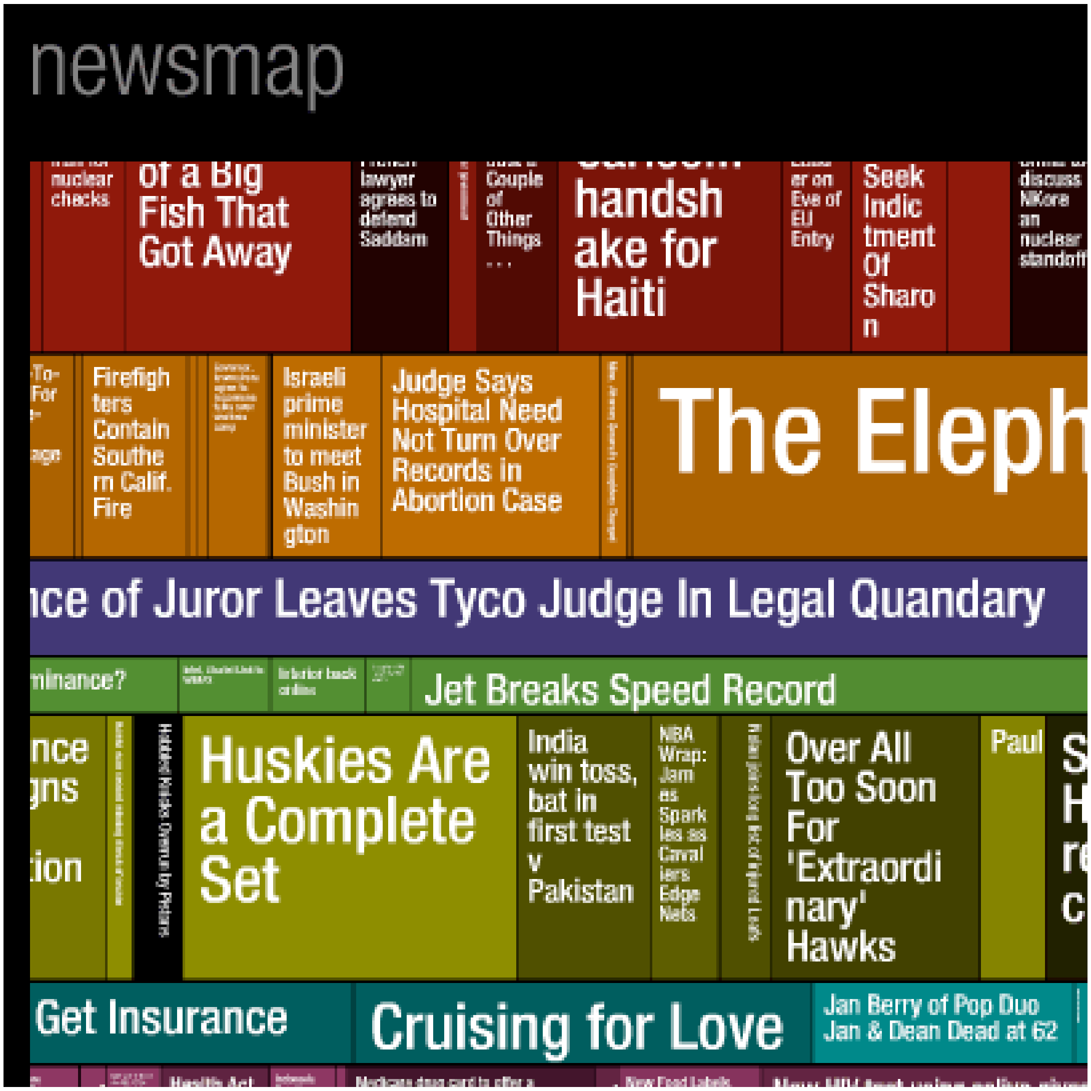
Figure 2: using Valence to visualize web site usage (35)

Marumushi - News Map: visual display of news

Marumushi's News Map provides a quick and effective way to browse the Google News news aggregated headlines. Headlines are tiled on the page in varying sizes and colors that correspond to different parameters. According to its creator (Marcos Weskamp)

Newsmap's aim was to visually organize the information in such a way that will reveal underlying patterns in news reports.

http://www.marumushi.com/apps/newsmap/newsmap.cfm



newsmap

nuclear checks

of a Big Fish That Got Away

lawyer agrees to defend Saddam

Couple of Other Things . . .

handsh ake for Haiti

er on Eve of EU Entry

Seek Indic tment Of Sharo n

discuss NKorea an nuclear standoff

Firefigh ters Contain Southe rn Calif. Fire

Israeli prime minister to meet Bush in Washin gton

Judge Says Hospital Need Not Turn Over Records in Abortion Case

The Eleph

nce of Juror Leaves Tyco Judge In Legal Quandary

minance?

Jet Breaks Speed Record

Hobbled Knicks Overrun by Pistons

Huskies Are a Complete Set

India win toss, bat in first test v Pakistan

NBA Wrap: Jam es Spark les as Caval iers Edge Nets

Over All Too Soon For 'Extraordi nary' Hawks

Paul

Get Insurance

Cruising for Love

Jan Berry of Pop Duo Jan & Dean Dead at 62

Health Act

Medicare drug card to offer a

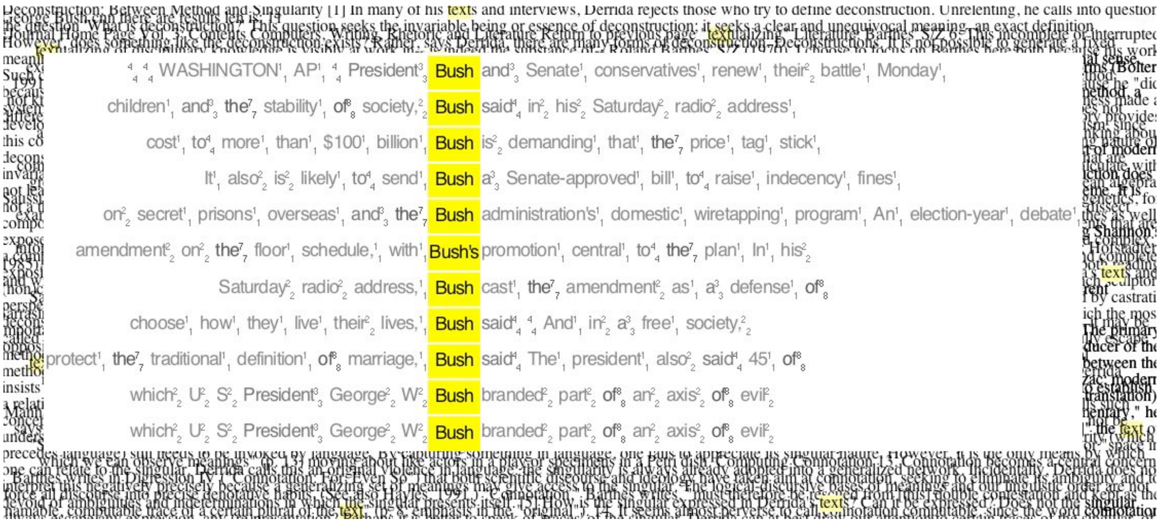
New Food Labels

figure 3: Marumushi News Map (36)

Ctrl-F reader

In my own “Ctrl-F reader” the user types a word, and chooses the content providers she is wishes to search. Then she is presented with lines extracted from texts that were added on that day. While the lines are presented in their entirety - the words become independent units. Every word ‘s visibility is proportional to its frequency in this day’s results, and is accompanied by information regarding its general frequency in all the results, and its relative frequency – in the results retrieved only from that specific space. Moreover every word becomes a link to a new search focusing on that word. If the user chooses to read the line in its context – she can click on the content provider’s logo, and she is immediately transferred to the page from this context was taken. If the user wishes to go further in the analysis of the language, she can click on the highlighted word and is presented with the next page, where the lines are broken into lonely words. There the words are presented according to their distance from the original search word, and data from previous searches in added.

www.ctrl-f.org



$^4_4$, $^4_4$, WASHINGTON$^1_1$, AP$^1_1$, $^4_4$, President$^3_3$ Bush and$^3_3$ Senate$^1_1$, conservatives$^1_1$, renew$^1_1$, their$^2_2$ battle$^1_1$, Monday$^1_1$,

children$^1_1$, and$^3_3$ the$^7_7$ stability$^1_1$, of$^8_8$ society,$^2_2$ Bush said$^4_4$ in$^2_2$ his$^2_2$ Saturday$^2_2$ radio$^2_2$ address$^1_1$,

cost$^1_1$, to$^4_4$ more$^1_1$, than$^1_1$, \$100$^1_1$, billion$^1_1$, Bush is$^2_2$ demanding$^1_1$, that$^1_1$, the$^7_7$ price$^1_1$, tag$^1_1$, stick$^1_1$,

It$^1_1$, also$^2_2$ is$^2_2$ likely$^1_1$, to$^4_4$ send$^1_1$, Bush a$^3_3$ Senate-approved$^1_1$, bill$^1_1$, to$^4_4$ raise$^1_1$, indecency$^1_1$, fines$^1_1$,

on$^2_2$ secret$^1_1$, prisons$^1_1$, overseas$^1_1$, and$^3_3$ the$^7_7$ Bush administration's$^1_1$, domestic$^1_1$, wiretapping$^1_1$, program$^1_1$, An$^1_1$, election-year$^1_1$, debate$^1_1$,

amendment$^2_2$ on$^2_2$ the$^7_7$ floor$^1_1$, schedule,$^1_1$ with$^1_1$, Bush's promotion$^1_1$, central$^1_1$, to$^4_4$ the$^7_7$ plan$^1_1$, In$^1_1$, his$^2_2$

Saturday$^2_2$ radio$^2_2$ address,$^1_1$ Bush cast$^1_1$, the$^7_7$ amendment$^2_2$ as$^1_1$, a$^3_3$ defense$^1_1$, of$^8_8$

choose$^1_1$, how$^1_1$, they$^1_1$, live$^1_1$, their$^2_2$ lives,$^1_1$ Bush said$^4_4$ $^4_4$ And$^1_1$, in$^2_2$ a$^3_3$ free$^1_1$, society,$^2_2$

protect$^1_1$, the$^7_7$ traditional$^1_1$, definition$^1_1$, of$^8_8$ marriage,$^1_1$ Bush said$^4_4$ The$^1_1$, president$^1_1$, also$^2_2$ said$^4_4$ 45$^1_1$, of$^8_8$

which$^2_2$ U$^2_2$ S$^2_2$ President$^3_3$ George$^2_2$ W$^2_2$ Bush branded$^2_2$ part$^2_2$ of$^8_8$ an$^2_2$ axis$^2_2$ of$^8_8$ evil$^2_2$

which$^2_2$ U$^2_2$ S$^2_2$ President$^3_3$ George$^2_2$ W$^2_2$ Bush branded$^2_2$ part$^2_2$ of$^8_8$ an$^2_2$ axis$^2_2$ of$^8_8$ evil$^2_2$

figure 4: Ctrl-F Reader "rips" George Bush from CNN

Conclusion:

In this essay I am trying to capture some of the changes in reading and research practices that are caused by the abundance of textual content in digital format. The existence of digitized text, and its abundance are simple facts, yet their influence is far reaching. I described the effects early electronic textual databases had on research practices, where only a fraction of the capacity of the machine was put to use. The machine was used in order to look for certain words, and extract whole lines containing those texts. Then I moved on to describe some more powerful machine algorithms search for frequent appearances of pairs of words, in order to search for synonyms. In this case the machine didn't "care" anymore for whole lines of text, but only for isolated words, and texts were further fragmented. In the last algorithmic reading of text I presented (Cyc, NGD), the machine moves on to analyze the syntactical composition of whole sentences, and tries to derive semantic meaning from them. While the context might seem reasonable, the "real world"context is at risk of being replaced by "machine logic" context.

Parallel to those algorithmic methods of reading I present some attempts at critical analytical approaches that try to deal with this content abundance. These methods rely on statistical tools which also remove the surveyed event out of its initial context.

Finally I describe some artistic projects, including my own, that propose new methods of visualizing textual content, when provided in digital format.

The overall conclusion from this essay is that there is a real need to develop new "reading aids" that will be able to assist us in processing the huge databases of textual digital content presented to us not only on the Internet, but also on any electronic textual database.

Thanks:

I wish to thank Zvi Lotker for innumerable insightful conversations, and Ami Asher for patient editing and correcting.

References:

1. www.Technorati.com, viewed on 20 April 2006.
2. http://www.nielsen-netratings.com/,,viewed on 20 April 2006.
3. http://2005.northernvoice.ca/node/131
4. http://en.wikipedia.org/wiki/Power_law
5. http://en.wikipedia.org/wiki/Google_bomb
6. http://geuzen.blogs.com/historiography/2005/11/lost_in_data.html
7. Saul, Richard Wurman Sume,David Leifer, Loring “Information Anxiety 2” Que, 2000

8. Ruhleder, Karen 'Pulling down' books vs. 'pulling up' files: textual databanks and the changing culture of classical scholarship". In "the Cultures of Computing. Susan Leigh Start (ed.), Blackwell Publishers 1995.
9. ibid. p. 186.
10. ibid. p. 188
11. ibid. p. 189
12. ibid. p. 188
13. ibid. p. 189
14. ibid. p. 189
15. Rapp, Reinhard "The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches", 2002
16. Ruge, G. (1992) "Experiments on Linguistically Based Term Associations" Information Processing & Management 28(3), 317–332.
17. Turney, Peter D. "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL"(2001)
18. Landauer, T.K., Dumais, S.T.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104 (1997) 211-240.
19. Fiske, John "Introduction to Communication Studies", Routledge, 2nd edition, 1990
20. ibid, p.136.
21. ibid p.137
22. ibid p.137

23. ibid p.139

24. ibid p.139

25. Stossel, Scott "The Man Who Counts The Killings" The Atlantic Monthly, (May 1997)

26. Florian Cramer, personal communication

27. Lakoff, George "don't think of an elephant! Know your values and frame the debate: the essential guide for progressives", Chelsea Green, 1st edition,(2004)

28. ibid p. xv

29. ibid p. 3

30. http://www.newscientisttech.com/article/mg18524846.100.html

31. Matuszek, Cynthia Witbrock, Michael Kahlert, Robert C. Cabral. John Schneider, Dave Shah, Purvesh Lenat, Doug. "Searching for Common Sense: Populating Cyc (tm) from the Web", (2005)

32. Cilibrasi, Rudi Vitanyi, Paul," Automatic Meaning Discovery Using Google",(2005)

33. ibid. p 19

34. www.complearn.org

35. http://acg.media.mit.edu/people/fry/valence/

36. http://www.marumushi.com/apps/newsmap/index.cfm

Bibliography:

New Media:

Espen, Aarseth J. “Cybertext. Perspectives on Ergodic Litterature”, The Johns Hopkins University Press (1997)

Brouwer, Joke Mulder, Arjen (ed.)(2003), Information is Alive,V2_Nai Publishers

Hayles, Katherine N. “My Mother Was a Computer”, The University of Chicago Press (2005)

Hayles, Katherine N. “Writing Machines”, Media Pamphlets MIT Press (2002)

Ong, J. Walter “Orality and Literacy” Rutledge (1982)

Rogers, Richard “Information Politics on the web”,MIT Press, (2004)

Saul, Richard Wurman Sume,David Leifer, Loring “Information Anxiety 2” Que, 2000

Susan Leigh Star (ed.), "the Cultures of Computing. Blackwell Publishers” (1995).

Wilson, Stephen, Information Arts, The MIT Press (2002) www.Technorati.com

http://www.nielsen-netratings.com/,,viewed on 20 April 2006.

http://2005.northernvoice.ca/node/131

Palisser, Nadia "Lost in Data"
http://geuzen.blogs.com/historiography/2005/11/lost_in_data.html

Computational Linguistics

Craven, Mark and DiPasqupo, Dan and Freitag, Dayne and McCallum, Andrew and Mitchell, Tom and Nigam, Kamal and Slattery, Sean "Learning to Extract Symbolic Knowledge from the World Wide Web"

Etzioni, Oren and Kok, Stanley and Soderland, Stephen, and Cafarella, Michael and Popescu, Ana-Maria, and Weld, Daniel S. and Downey, Doug and Shaked, Tal and Yates, Alexander "Web-Scale Information Extraction in KnowItAll (Preliminary Results)"

French, R. M. and Labiouse, C. “Why co-occurrence information alone is not sufficient to answer subcognitive questions.” Journal of Theoretical and Experimental Artificial Intelligence, 13(4), 419-429. (2001)

Lin, Dekang and Pantel Patrick “Concept Discovery from Text”

Rapp, Reinhard “The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches”, (2002)

Turney, Peter D. “Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL”,(2001)

Turney, Peter D. “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification Reviews”

Wettler, Manfred, Rapp, Reinhard “ Computation of Word Associations Based on the Co-Occurrence of Words in Large Corpora” (1993)

CL Intro text: http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

Communication Studies:

Fiske, John “Introduction to Communication Studies”, Routledge, 2nd edition, (1990)

Lakoff, George “don’t think of an elephant! Know your values and frame the debate: the essential guide for progressives”, Chelsea Green, 1st edition,(2004)

Stossel, Scott “The Man Who Counts The Killings” The Atlantic Monthly, (May 1997)

Computer Science:

Cilibrasi, Rudi Vitanyi, Paul “Automatic Meaning Discovery Using Google” (2005)

Matuszek, Cynthia Witbrock, Michael Kahlert, Robert C. Cabral. John Schneider, Dave Shah, Purvesh Lenat, Doug. “Searching for Common Sense: Populating Cyc (tm) from the Web”, (2005)

CommonSense Computing : http://xnet.media.mit.edu/

Fiction:

Borges, Jorge Luis, Fictions, Penguin Classics (1944)